บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เอกสารหนังสือหรือสิ่งพิมพ์ต่างๆที่เราพบเห็นรอบๆตัวในปัจจุบัน จะจัดพิมพ์บนกระดาษเป็นส่วนใหญ่ บางเล่มเอกสารก็มีขนาดใหญ่น้ำหนักมากไม่สะดวกในการพกพา หรือบางเอกสารมีเนื้อหามาก การค้นหาสิ่งที่เราสนใจในเอกสารนั้นก็ยากในการค้นหรือต้องใช้เวลามาก ในปัจจุบันเทคโนโลยีและคอมพิวเตอร์ได้พัฒนาไปอย่างรวดเร็ว และเข้ามาเป็นส่วนหนึ่งในการดำรงชีวิตของคนในปัจจุบัน ทำให้การทำงานหลายๆอย่างมีความสะดวกสบายและรวดเร็วมากขึ้น ในงานเอกสารและสิ่งพิมพ์เช่นกัน เอกสารไม่ต้องบันทึกไว้ในกระดาษไม่ต้องมีขนาดใหญ่โต ข้อมูลเนื้อหาเหล่านั้นจะถูกบันทึกไว้ในสื่ออิเล็กทรอนิกส์ ซึ่งมีขนาดกระทัดรัดและมีความสามารถในการบันทึกข้อมูลได้มาก ทั้งการเรียกค้นหาข้อมูลด้วยระบบคอมพิวเตอร์ก็ทำได้สะดวกรวดเร็วกว่าการค้นหาโดยการอ่านทางสายตา

เอกสารพีดีเอฟ (PDF ย่อมาจาก Portable Document Format) คือ เอกสารอิเล็กทรอนิกส์ ที่มีความยืดหยุ่นเป็นอิสระไม่ขึ้นกับระบบปฏิบัติการของโปรแกรมสำเร็จประยุกต์ซอฟต์แวร์ ฮาร์ดแวร์ ที่ใช้สร้างเอกสาร[1] เช่น สมมุติว่า บริษัทผ้าล้านนาเป็นบริษัทที่ทำกิจการในการออกแบบ ผลิต และส่งออกผ้า ใช้โปรแกรมสำเร็จประยุกต์ “ลายกนก” บนระบบปฏิบัติการแมคอินทอช(Macintoch) ในการออกแบบลายผ้า ผู้ใช้ต้องส่งลายผ้าไปให้บริษัทลูกค้าเพื่อดูตัวอย่างลายผ้า บริษัทผ้าล้านนาก็จะสร้างแบบลายพิมพ์นั้นให้อยู่ในรูปแบบเอกสารพีดีเอฟ และส่งเป็นไปรษณีย์อิเล็กทรอนิกส์ไปให้กับบริษัทลูกค้า ทางบริษัทลูกค้าไม่มีโปรแกรมสำเร็จประยุกต์ “ลายกนก” ไม่ได้ใช้ระบบปฏิบัติการแมคอินทอชแต่ใช้ระบบปฏิบัติการอื่น ก็สามารถที่จะอ่านแบบลายผ้าได้เหมือนกับแบบลายผ้าต้นฉบับที่บริษัทผ้าล้านนาออกแบบไว้

มีโปรแกรมสำเร็จประยุกต์ที่ [illegible] สามารถใช้ดูเอกสารพีดีเอฟ แต่ที่ได้รับความนิยมมากที่สุด ได้แก่ อะโครแบทรีดเดอร์(Acrobat Reader®) และ อะโครแบทเอ็กเซนจ์(Acrobat Exchange®) ซึ่งเป็นของบริษัทอะโดบี Adobe Inc เนื่องจาก มีคุณสมบัติที่อำนวยความสะดวก

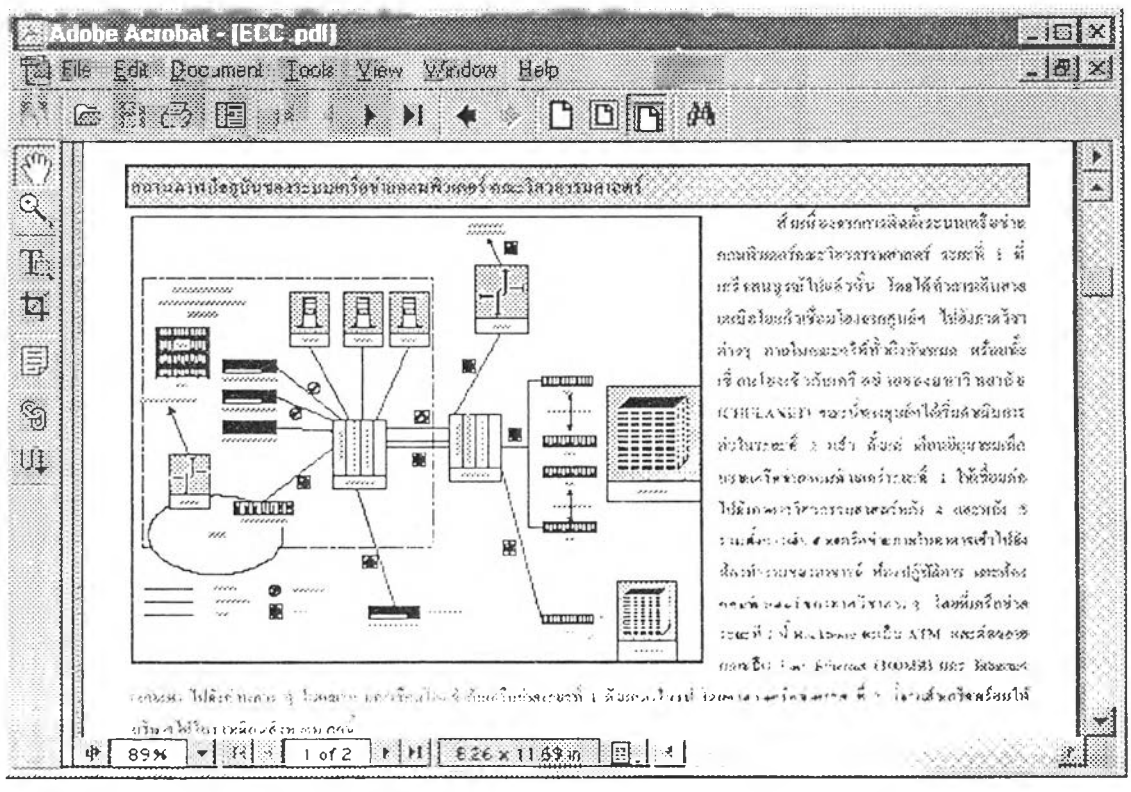
---

[1] Adobe Systems Incorporated., Portable Document Format Reference Manual. Adobe Systems, 1999, p. 19

แก่ผู้ใช้หลายอย่าง อาทิเช่น ความสามารถในการเชื่อมโยงแต่ละหัวข้อในเอกสาร ความสามารถในการแสดงเอกสารแบบย่อขยายขนาดเอกสาร ความสามารถในการแสดงเสียงหรือภาพเคลื่อนไหวในเอกสารความสามารถในการพิมพ์ และอื่นๆอีกมากมาย

อะโครแบทรีดเดอร์ เป็นโปรแกรมสำเร็จประยุกต์ที่ใช้ในการดูเอกสารพีดีเอฟที่แจกฟรี ผู้ใช้ไม่ต้องเสียค่าใช้จ่ายในการนำมาใช้งาน แต่มีข้อจำกัดในการใช้เอกสารคืออนุญาตให้ผู้ใช้สามารถทำได้เฉพาะ การดูเอกสาร การค้น การพิมพ์ และอื่นๆที่ไม่ใช้การแก้ไขเอกสาร อะโครแบทเอ็กเซนจ์มีคุณสมบัติของอะโครแบทรีดเดอร์อยู่ครบถ้วนและยังอนุญาตให้แก้ไขเอกสารได้แต่ผู้ใช้ต้องเสียค่าใช้จ่ายในการนำมาใช้ เนื่องจากเป็นผลิตภัณฑ์ที่มีไว้จำหน่าย

ผู้ใช้งานและนักวิชาการในประเทศไทย ก็ให้ความสนใจและมีความนิยมเพิ่มขึ้นที่จะสร้างเอกสารให้อยู่ในรูปแบบเอกสารพีดีเอฟ โดยจะพบได้จาก บทความ สิ่งพิมพ์ต่างๆในปัจจุบันจะอยู่ในรูปแบบเอกสารพีดีเอฟมีจำนวนเพิ่มมากขึ้น (ดูรูปที่ 1 ประกอบ) อาทิเช่น เอกสารของศูนย์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาฯ (www.eng.chula.ac.th) เอกสารบทความของนักวิชาการบางท่าน (www.pharm.chula.ac.th/surachai) เอกสารที่ใช้ประกอบในการเรียนการสอน (www.cpc.ku.ac.th/~semina) สิ่งพิมพ์นิตยสารต่างๆ (www.mcot.or.th/v_bookworld1) นิตยสารขวัญเรือน (www.thaimag.com/kwanruen) และยังมีอื่นๆอีกมากมาย

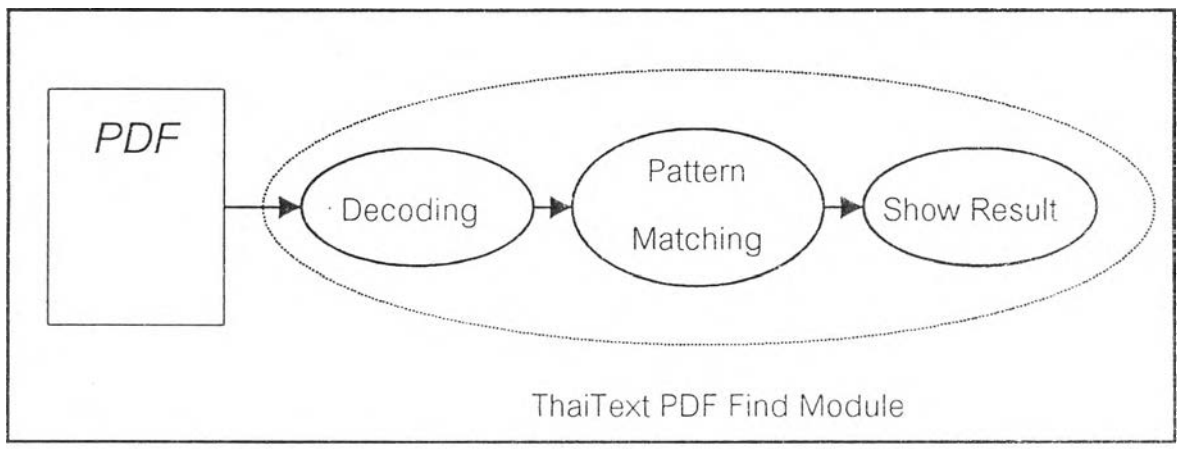


รูปที่ 1 เอกสารพีดีเอฟศูนย์คอมพิวเตอร์คณะวิศวกรรมศาสตร์ จุฬาฯ

แต่พบว่า เอกสารพีดีเอฟที่ใช้ภาษาไทยยังมีข้อบกพร่องอยู่มาก ปัญหาหนึ่งที่พบก็คือ เครื่องมือการค้นในโปรแกรมแสดงเอกสารพีดีเอฟไม่สามารถค้นข้อความไทยได้อย่างถูกต้อง ทำให้เอกสารพีดีเอฟที่ใช้แบบตัวอักษรและอักขระไทยใช้ดูได้เพียงอย่างเดียว แต่เอกสารไม่สามารถที่จะค้นได้ ซึ่งการค้นเป็นคุณสมบัติหนึ่งที่สำคัญของเอกสารอิเล็กทรอนิกส์ ทำให้ใช้เอกสารพีดีเอฟที่ใช้ภาษาไทยได้ไม่เต็มความสามารถ ปัญหาเหล่านี้เกิดจาก

1. ตัวอักษรภาษาไทยในเอกสารพีดีเอฟ มีการเข้ารหัสตัวอักษรที่ไม่เป็นไปตามมาตรฐานตามข้อกำหนดของเอกสารพีดีเอฟ
2. กระบวนการในการค้นข้อความของเครื่องมือการค้นในโปรแกรมแสดงเอกสารพีดีเอฟไม่ได้ถูกออกแบบมา เพื่อให้สนับสนุนกับการเข้ารหัสตัวอักษรที่ตัวอักษรภาษาไทยใช้ จึงไม่เข้าใจการเข้ารหัสตัวอักษรภาษาไทย ทำให้ค้นข้อความไทยได้ไม่ถูกต้อง

จากที่ได้กล่าวมาข้างต้น เครื่องมือการค้นที่มีอยู่ไม่สามารถค้นข้อความไทยได้ วิธีการในการแก้ปัญหา ผู้เขียนจึงจะสร้างกระบวนการค้นข้อความที่เข้าใจการเข้ารหัสตัวอักษรของชุดแบบภาษาไทยที่ใช้ในเอกสารพีดีเอฟ โดยจะมีแนวคิดดังนี้ (ดูรูปที่ 2 ประกอบ)



รูปที่ 2 แนวคิดกระบวนการค้นข้อความไทยในเอกสารพีดีเอฟ

ขั้นตอนในการค้นข้อความไทยในเอกสารพีดีเอฟ จะประกอบด้วยขั้นตอนต่างๆ ดังต่อไปนี้

1. การถอดรหัสข้อความไทยในเอกสารพีดีเอฟ (Decoding)

ตัวอักษรทุกตัวในเอกสารพีดีเอฟจะประกอบด้วย รหัสตัวอักษรและข้อมูลการเข้ารหัสตัวอักษร เมื่อผู้ใช้เปิดอ่านเอกสารพีดีเอฟ โปรแกรมในการแสดงเอกสารพีดีเอฟจะนำรหัสตัวอักษรและข้อมูลการเข้ารหัสไปใช้ในการแสดงแบบอักษร กระบวนการค้นข้อความไทยในเอกสารพีดีเอฟก็เช่นกัน จะต้องทำการถอดรหัสตัวอักษรก่อน เพื่อให้ได้รหัสตัวอักษรที่ตรงกับความหมายของการ

เข้ารหัส แต่เนื่องจากข้อความภาษาไทยที่ใช้ในเอกสารพีดีเอฟมีการเข้ารหัสตัวอักษรที่ไม่เป็นไปตามข้อกำหนดมาตรฐานการเข้ารหัสที่เอกสารพีดีเอฟกำหนด ดังนั้นในการถอดรหัสข้อความไทยในเอกสารพีดีเอฟจะต้องมีกระบวนการที่สามารถจะเข้าใจการเข้ารหัสที่ไม่เป็นมาตรฐานต่างๆ เหล่านั้น

2. การค้นข้อความโดยวิธีการเปรียบเทียบสายอักขระ (Pattern Matching)

เมื่อได้รหัสข้อความจากขั้นตอนการถอดรหัสแล้ว ขั้นตอนถัดมาจะทำการค้นข้อความโดยวิธีการเปรียบเทียบอักขระ การเปรียบเทียบอักขระจะเป็นการเปรียบเทียบกลุ่มสายอักขระ 2 กลุ่มจากซ้ายไปขวาที่ละ1คู่อักขระ แต่เนื่องจากในขั้นตอนการถอดรหัสข้อความที่อยู่ในเอกสารพีดีเอฟจะถูกแยกออกเป็นได้หลายส่วน ซึ่งกระบวนการเปรียบเทียบสายอักขระโดยทั่วไปไม่สามารถค้นได้ถูกต้อง ดังนั้นในการค้นข้อความไทยในเอกสารพีดีเอฟโดยวิธีการเปรียบเทียบสายอักขระ จะต้องมีกระบวนการเพิ่มเติมเพื่อตรวจสอบการค้นข้อความที่ถูกแยกเป็นข้อความย่อยๆ

ตัวอย่างเช่น ข้อความที่ได้จากการถอดรหัส คือ "ธนาคาร" + "กรุ" + "งเทพ" + "จำกัด"

ข้อความที่ต้องการค้น คือ "กรุงเทพ"

ถ้าใช้วิธีการค้นโดยทั่วไป จะได้ผลลัพธ์ว่าไม่พบข้อความที่ต้องการค้น

3. การแสดงผลการค้นข้อความ (Show Result)

เมื่อทำการค้นข้ความไทยในเอสารพีดีเอฟแล้ว จะทำการตอบสนองกับผู้ใช้ให้ผู้ใช้ทราบว่าพบข้ความหรือไม่ ถ้าพบข้อความจะทำการแสดงสีทึบทับข้อความที่ค้นพบ

วัตถุประสงค์

เพื่อออกแบบและพัฒนาส่วนจำเพาะการค้นข้อความไทยในเอกสารพีดีเอฟ

ขอบเขตการวิจัย

1. เอกสารที่จะค้นต้องเป็นเอกสารที่มีการเข้ารหัสตัวอักษรตรงตามมาตรฐาน มอก.620 หรือ มีการเข้ารหัสที่รหัสตัวอักษรลดลงด้วยค่าที่คงที่ หรือ ชื่อตัวอักษรในการเข้ารหัสไม่ถูกเปลี่ยนชื่อไป
2. ส่วนจำเพาะนี้ทำงานบนระบบปฏิบัติการวินโดวส์ 32 บิต ร่วมกับอะโครแบทรีดเดอร์หรืออะโครแบทเอ็กเซนจ์

3. ส่วนจำเพาะมีคุณสมบัติ ดังนี้ สามารถค้นข้อความภาษาไทยหรือภาษาอังกฤษ โดยค้นไปข้างหน้า ค้นย้อนหลังกลับไป คัดข้อความออกจากเอกสาร นำมาเก็บไว้เป็นแฟ้มข้อความ หรือ ปะข้อความเป็นหมายเหตุติดเข้าไปไว้กับเอกสารพีดีเอฟ

ขั้นตอนการวิจัย

1. ศึกษาปัญหาและแนวทางการแก้ไขที่เป็นไปได้
2. วิเคราะห์และวางแผนขั้นตอนการทำงาน
3. ออกแบบและพัฒนาส่วนจำเพาะ โดยทำการพัฒนาดังนี้
   3.1 พัฒนาส่วนการคัดข้อความและข้อมูลแบบอักษรในเอกสารพีดีเอฟ
   3.2 พัฒนาส่วนการนำข้อความมาเก็บไว้เป็นแฟ้มข้อความ
   3.3 พัฒนาส่วนการวิเคราะห์การเข้ารหัสตัวอักษร
   3.4 พัฒนาส่วนการค้นข้อความ
4. ทดสอบโปรแกรมและแก้ไขส่วนที่บกพร่อง
5. สรุปผลการวิจัย และจัดทำรายงานวิทยานิพนธ์

ผลที่คาดว่าจะได้รับ

สามารถอำนวยความสะดวกผู้ใช้ ในการค้นข้อความที่ต้องการในเอกสารพีดีเอฟ